Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 5665–5672
Marseille, 20-25 June 2022

Building a Synthetic Biomedical Research Article Citation Linkage Corpus

Sudipta Singha Roy and Robert E. Mercer
The University of Western Ontario
London, Ontario, Canada
ssinghar@uwo.ca, mercer@csd.uwo.ca

Abstract

Citations are frequently used in publications to support the presented results and to demonstrate the previous discoveries while also assisting the reader in following the chronological progression of information through publications. In scientific publications, a citation refers to the referenced document, but it makes no mention of the exact span of text that is being referred to. Connecting the citation to this span of text is called *citation linkage*. In this paper, to find these citation linkages in biomedical research publications using deep learning, we provide a synthetic silver standard corpus as well as the method to build this corpus. The motivation for building this corpus is to provide a training set for deep learning models that will locate the text spans in a reference article, given a citing statement, based on semantic similarity. This corpus is composed of sentence pairs, where one sentence in each pair is the citing statement and the other one is a candidate cited statement from the referenced paper. The corpus is annotated using an unsupervised sentence embedding method. The effectiveness of this silver standard corpus for training citation linkage models is validated against a human-annotated gold standard corpus.

Keywords: citation linkage, semantic similarity, automatic corpus annotation

1. Introduction

There are a variety of formats, writing styles, and purposes for different types of written documents. It is possible for a research article to reflect a current trend in the field of study, a new invention, or a novel approach to solving a specific problem. During the process of writing a research paper, the author examines past studies that are either important in solving the topic at hand or have impacted the author's current research paper ideas. Using a *citation* is the process to refer to another article in the current research article (Houngbo, 2017). In this way, citations serve as bridges between different research papers. Citations free up the authors' time by removing the need to repeatedly write the same thing. While doing so, it provides readers with some context for the issues being discussed in the body of the piece.

The concept of citation indexing was first introduced in 1964 by Garfield (1972) where indexes contain the entirety of the references in a research document. Since then, various analyses of citing have been presented (e.g., (Ritchie et al., 2008)). In biochemistry and physics research papers, Garzone and Mercer (2000) presented a method for determining the objectives of different citations. Furthermore, citation aids in the tracking of logical argumentation throughout multiple research articles (Mercer, 2016). Citation is commonly used to maintain the trail of scientific research argumentation across different scientific papers (Palau and Moens, 2009) and to summarise these documents (Radev et al., 2000).

When writing scientific research publications, citations are used when referring to a source of inspiration for a cited idea. In the case of experimental biomedical research, only a small portion of the referred material, which can be from the methodological, result, or any other sections of the cited document, is often relevant. Applications like the ones listed above would benefit from being able to extract just that relevant portion of the cited document's text. In addition, readers would not have to read an entire referenced document in order to locate the mentioned text span.

The citation linkage task for biomedical literature is a complex process: a chemical compound can be presented in multiple ways; the reactions between different drugs, chemical components, and genes can be described in very different manners; and for research articles from different sub-domains of this field, this information can be represented in different ways. Furthermore, not a lot of resources are available for deep learning this task as annotating a large corpus takes a lot of time and the annotators require domain-knowledge. At the same time, deep learning based models are data hungry and require a lot of annotated data for such task. A few corpora for the citation linkage task are currently available, but almost all are for the domain of computational linguistics research articles, not for biomedical research literature (Li et al., 2019).

The objective of this paper is to present a method for generating a synthetic silver standard corpus for the citation linkage task for biomedical research articles and to introduce a corpus containing 74,568 sentence pairs to the research community. This corpus contains sentence pairs that are tagged as being semantically similar or not. However, since we are using semantic similarity as a proxy for citation linkage, the corpus is intended to train models which view the citation linkage task as a textual semantic similarity measurement task in the same way as Li et al. (2019). We call this corpus a synthetic corpus as the dataset is annotated by unsu-

Table 1: Sample citations and the intended reference sentences that correspond (from: Houngbo (2017))

Example 1	Citing Statement	Formalin fixation, the most often used fixative in histology, has various advantages, including ease of tissue manipulation, optimal histological quality, long-term preservation capability, and widespread availability at a reasonable cost. (Huijsmans et al., 2010)
	Cited Statement	The advantages of using formalin fixation are simplicity of tissue handling, the ability to store wet material for an extended period of time, and its inexpensive cost. (Kayser et al., 1988)
Example 2	Citing Statement	DNA samples are frequently harmed by exposure to excessively acidic environment. (Wang et al., 2009)
	Cited Statement	DNA is fairly stable in mildly acidic solutions, although the beta glycosidic link in the purine bases is hydrolyzed at around pH4. (Bonin et al., 2003)
Example 3	Citing Statement	Different PCR buffer systems and/or Taq polymerases may produce variable results in real time PCR. (Huijsmans et al., 2010)
	Cited Statement	There is a significant disparity between the outcomes obtained using the various DNA polymerase-buffer solutions. (Wolffs et al., 2004)

pervised sentence embedding models, not by humans. And finally, the effectiveness of this dataset is assessed by testing some linear and tree-structured neural network models, which are trained with this silver corpus, on a human annotated gold corpus. The following is how the remainder of the paper is organised: The citation linkage task is discussed in Section 2 while Section 3 provides some relevant research which tackles the citation linkage task by means of assessing textual semantic relatedness between the citing and cited text spans. Data collection, data cleaning, and the automatic silver corpus creation steps are discussed in Section 4. In Section 5, the assessment of the effectiveness of this corpus is analyzed. Finally, this paper concludes with a brief summary of this work along with some directions for future research.

2. Citation Linkage

Citations create a semantic connection between the articles that are citing and the manuscripts that are being cited. While writing a research article, the authors use reference articles to support their findings and hypotheses. At the same time, they try to acknowledge the findings of the other researchers. Mentioning others' works is also important to show the significance and improvements brought by the authors with their current work.

A citation inside a research article refers to a section of the reference paper known as the *citation context* (Houngbo, 2017). An idea or issue addressed in the referenced work is often the focus of this citation context. The citation intends to give some insight about the apposite background information to the reader so the concept of the ongoing paper becomes more understandable to them. It is possible to identify the methods, instruments, or discoveries and hypotheses in a cited publication by looking at the citation context. An author may adapt the method mentioned in the citing paper or modify it to some extent so that the performance improves or becomes compatible to the domain where he/she wants to deploy that method. Moreover, the author may conduct some experiments based upon the hypothesis of the cited paper. References to those used methods and hypotheses help the readers to easily grasp the ideas presented in the ongoing paper.

Citations, on the other hand, do not specify which part of the referenced article is being alluded to; rather, they simply state the title of the cited piece. As a result, if a reader is interested in learning more about the issue, he or she has to study the entire cited document. Readers, on the other hand, like research articles that provide them with specifics on the findings that were made during the study with clear and specific background knowledge. This necessitates a clear understanding of the influences that have shaped this work.

A few examples of citation sentences and their related reference sentences in the cited publication are shown in Table 1. In Example 1 a paraphrase of the cit+ed sentence is given which incorporates common words in a different sequence in the citing sentence. The term "pH4" is replaced by "excessively acidic environment" in the second example. It is necessary to map the pH scaling to the acidic situation to connect these two ideas. The citation sentence in Example 3 interprets the target sentence's information. It is obvious from these examples that accurate mapping between sentences and words is necessary for creating the relationship between the citing and referenced sentences. This paper presents a synthetic silver standard corpus for training models to solve the citation linkage task for biomedical research articles by means of measuring semantic relatedness between the citing and candidate cited statements. Usually, the citation context can comprise from one single sentence to multiple paragraphs. However, models trained on this corpus can link related sentences from the cited paper given the citing sentences from the ongoing paper. This corpus comes with sentence pairs where one sentence in each pair is the citing statement and another sentence in the pair is the candidate cited statement from the reference paper. The sentence pairs in this corpus are labeled with either

0 or 1, where 1 indicates the sentences in the pair are semantically similar and 0 denotes dissimilarity.

3. Related Works

There has been a significant amount of work done to analyse citations in scientific research publications as a result of growing interest in citations (Garfield, 1972; Garzone and Mercer, 2000). One approach is using citation analysis to figure out which area (such as the abstract, introduction, methodological description, result analysis and discussion of the findings) of a cited article is being referenced by a certain citation sentence. An exact citation span cannot be determined using this this type of analysis.

To help with the citation linkage task, the CL-SciSumm Shared Task is examining three different aspects: finding the text span in the referenced paper that best captures each citation sentence (a "citance"); identifying the discourse facet of each cited text span; and the reference paper's summarization using text spans referenced by several citances. The last two tasks go beyond the scope of the current paper. Text granularity considered in the first task are complete sentences, fragments of sentences, and up to five sequential sentences. In this study, while creating the corpus, we considered single sentences as the cited text span. A corpus of computational linguistics research papers is used in the CL-SciSumm Shared Task.

For the CL-SciSumm-17 shared task, Li et al. (2017) used Jaccard similarity and inverse document frequency to assess which sentence pairs in citing and cited sources were linked to one another. Li et al. (2018) computed the cosine similarity between sentence vectors. These sentence vectors were the concatenations of the corresponding words' 200 dimensional vectors computed from word2vec (Mikolov et al., 2013). In this work, they applied a convolutional neural network over these sentence representations for generating better feature representations. Gidiotis et al. (2020) fine-tuned BERT for generating sentence representations for the very same task. Umapathy et al. (2020) used the Rapid Automated Keyword Extraction Algorithm (Rose et al., 2010) for detecting key-phrase similarity and a BERT-based model for detecting citation text span.

Regrettably, just a few works in the biomedical field have attempted this citation linking endeavour. And that's why only one gold standard human annotated corpus is available for this task in the biomedical field. In 2017, Houngbo and Mercer (2017) created a small expert-annotated corpus consisting of sentence pairs from the biomedical area and used different traditional machine learning algorithms for textual matching operations to establish a framework for the citation linkage task.

4. Corpus Creation

In the biomedical domain, the only human annotated gold standard corpus available is from Houngbo and Mercer (2017)'s work. This corpus covers texts and citations only from the methodological sections from the biomedical research articles. The citation text span in this corpus is limited to only one sentence. So, the models trained on this corpus are designed for measuring semantics of the sentence pairs, though the citation text spans in scientific research papers may cover one or multiple sentences and from different portions of the articles. The corpus is annotated by experts with proper domain knowledge and contains 3857 sentence pairs with 23 citing statements. The sentence pairs are annotated on a scale of 1 to 5 ((minimum to maximum similarity between the citing and candidate cited statement) and 0 (no similarity between citing and candidate cited sentence).

The major problem while working with this corpus is the highly imbalanced proportion between positive and negative samples. Out of these 3857 samples present in this corpus, only 81 samples are annotated with similarity score 4 and 5. That's why models trained with this corpus become highly biased towards the negative outcome. On the other hand, annotating a corpus with a sufficient number of samples which is balanced in proportion of the positive and negative samples is a very time consuming process and demands expert domain knowledge. And without such a dataset, it is tough to train data hungry deep learning models for the citation linkage task in the biomedical domain. To overcome these shortcomings, we present our synthetic corpus of 74,568 sentence pairs from 2,736 citing and 138 cited papers covering 3 biomedical sub-domains: chemical biology, biochemistry and cell biology. We call this corpus synthetic as no human supervision is used for data annotation. Rather, the unsupervised sentence embedding model Sent2Vec (Pagliardini et al., 2018) is used to serve this purpose. For assessing the effectiveness of this synthetic corpus, models are trained with this corpus, but validated and tested against the gold standard corpus of Houngbo and Mercer (2017)'s work. However, the scoring factor of this gold standard corpus is modified for our work. Similarity scores of the samples with score 0 to 3 are replaced by 0 and samples annotated with similarity score 4 and 5 are labelled with 1. We chose Sent2Vec for creating the silver standard synthetic corpus because of it's ability to work with out of vocabulary words and doesn't require any pre-trained word embeddings (Pagliardini et al., 2018).

The overall corpus creation process is described here in three steps: i) data collection, ii) data cleaning, and iii) data annotation.

4.1. Data Collection

Sent2Vec, like all other unsupervised models, demands a large amount of training data. That's why for training the model, 4,843,756 sentences from 28,310 research documents are accumulated. These documents from more than 90 different fields of biomedicine are extracted from BioMed Central.

Table 2: Regex for detecting distinct patterns in the data

Purpose	Regex command
Capture equations	[\{\{\{(\[\s]*([^\s\{\{\{(\[\}\)\}\]%*+/:<=>\-\^|~¬±·×÷ϴθ— ←→↔↦≒⇔∀∇∈∉∈∌∏∑∓∘∨∝∧∩∪∮≁≤≥≰⊂⊄⊆⊇⊈⊑⊕ ⊙⊣⊤⊥∪·|⩟▷○∘⟶⫫]+(\s*[\{\{\{(\[]+\s*)+)?(∃|∫|log|ln)?\s* ([^\s\{\{\{(\[\}\)\}\]]+|(\d+[\.,])+\d+)?[\)\)\)\}\]]*([\{\{\{(\[\s]*[%*+/:< =>\^|~¬±·×÷ϴθ— ←→↔↦≒⇔∀∇∈∉∈∌∏∑∓∘∨∝∧∩∪∮≁≤≥≰⊂⊄⊆⊇⊈⊑⊕ ⊙⊣⊤⊥∪·|⩟▷○∘⟶⫫]+\s*- ?[\{\{\{(\[\s]*(∃|∫|log|ln)?\s*([^\s\{\{\{(\[\}\)\}\]]+|(\d+[\.,])+\d+) \s*([\{\{\{(\[\s]*[^\s\{\{\{(\[\}\)\}\]%*+/:<=>\^|~¬±·×÷ϴθ— ←→↔↦≒⇔∀∇∈∉∈∌∏∑∓∘∨∝∧∩∪∮≁≤≥≰⊂⊄⊆⊇⊈⊑⊕ ⊙⊣⊤⊥∪·|⩟▷○∘⟶⫫]+|(\d+[\.,])+\d+|-)?[\s\)\)\)\}\]]*)+
Capture numbers with no prior symbol	(^\d+| \K\d+)((\.\d+)|\^\d+|e[\-+]?\d+(\.\d+)?)?(?![\-\w])
Capture numbers with prior symbol	(?<!(\w|\d))[\-+]\d+((\.\d+)|\^\d+|e[\-+]?\d+(\.\d+)?)?
Capture citations	\[\d+([,\-]\d+)*\]
Capture chemical names ending with ∝ and β	-\s*∝|-\s* β

For the purpose of creating the sentence pairs with citing and cited sentences, 138 articles from the fields of cell biology, biochemistry, and chemical biology were chosen at random from a pool of these 28,310 research papers and these papers are considered as the cited reference papers. A total of 2,736 citing papers (cite at least one of the papers from these 138 reference papers) are collected manually and from them, only the relevant citing statements are extracted.

4.2. Data Cleaning

As the research articles are accumulated from different biomedical sub-domains, they come with a variety of writing formats and representations. Furthermore, the same equations may be represented in different ways with different symbols and variable names. That's why, to avoid confusion, all of the equations in these cited and citing papers are replaced with "$<$ *equ* $>$". All of the isolated numbers are also replaced by "$<$ *num* $>$". However, if any number comes as a part of any chemical compound name, it is preserved without any modification. The documents contain some symbols which have no importance in terms of representing the semantics neither at the sentence level nor at the document level. Such symbols are identified and deleted from the data. Citation indexes like "[xx]" are also deleted as they have no semantic value. Some Greek letters have different usages in different scenarios. For instance, α has no importance in terms of semantics when it is used as a variable in an equation, but, when it comes as a part of a chemical name, like "α-carbon", it differentiates the chemical from other variants. That's why when such Greek letters appear as a part of equations, they are kept untouched and the whole equation is replaced with "$<$ *equ* $>$". But, when these Greek letters come as a part of chemical names, they are replaced with their written form (e.g., α in "α-carbon" is replaced by alpha). Finally, symbols which are represented in multiple ways are replaced with their corresponding common format of representations and then all the data are lower-cased. Table 2 shows the regex commands used for the data cleaning step. Finally, all the unnecessary symbols are deleted.

4.3. Data Annotation

Following the cleaning of the sentences, the unsupervised sentence embedding model Sent2Vec is trained. This step is necessary in order to properly annotate the pairs of citing and candidate cited sentences. That is why Sent2Vec is trained on the data using a variety of parameter settings. Table 3 illustrates different hyper-parameter settings. Hyper-parameter values for the optimal sentence embedding are indicated, as well.

In order to produce a sentence pair for each sentence in the cited article, after the data has been cleaned, sentence pairs are formed in which one sentence is taken from the cited article and the other is taken from the citation. A total of 522,398 sentence pairings are gen-

Table 3: Hyper-parameter settings used for training Sent2Vec. The selected parameter values are marked as bold.

Hyper-parameters	Values
Embedding Dimension	700/600/**500**/400/300/200
Iterations	20/15/**10**/5
Window Size	**20**/10
Learning Rate	**0.2**/0.1/0.05/0.01
Negative Samples	**10**
Loss Function	softmax/ Hierarchical softmax/ **Negative sampling**
Sampling Threshold	**0.0001**

erated in this step.

The Sent2Vec model is then used to generate the vector representations of individual sentences from each pair and after that, cosine similarity between sentence vectors for citing and candidate cited statements in each pair is computed. Performance is evaluated against the gold standard validation set from Houngbo and Mercer (2017)'s work for varied cutoff cosine similarity values. This validation set consists of 800 sentence pairs with 20 randomly chosen positive samples. Samples with cosine similarity score more than the cutoff are tagged with similarity score 1 (indicates the citing and the candidate cited statements are semantically similar) and 0 otherwise (there is no similarity between the citing and the candidate cited sentences). This cutoff value is determined by looking at the Balanced Accuracy, Matthews correlation coefficient (MCC), and F1 score metrics over the validation dataset. Sentence vectors with 500-dimensional representations and a cutoff value of 0.57 produce the best results.

However, after this approach it is found that the vast majority of these 522,398 sentence pairs have annotation value 0. Any model will be biased towards the negative outcome, if it is trained with this corpus. Because of this, 74,568 samples are selected from these pairs to ensure that the positive and negative samples are evenly distributed. For this selection process, all the positive samples (annotated with similarity value 1) are retained, while for each citing statement, n negative samples are chosen randomly where for that citing sentence n positive samples are found. Thus, this evenly distributed silver standard corpus with 74,568 is generated. The whole corpus creation process is portrayed in Fig 1.

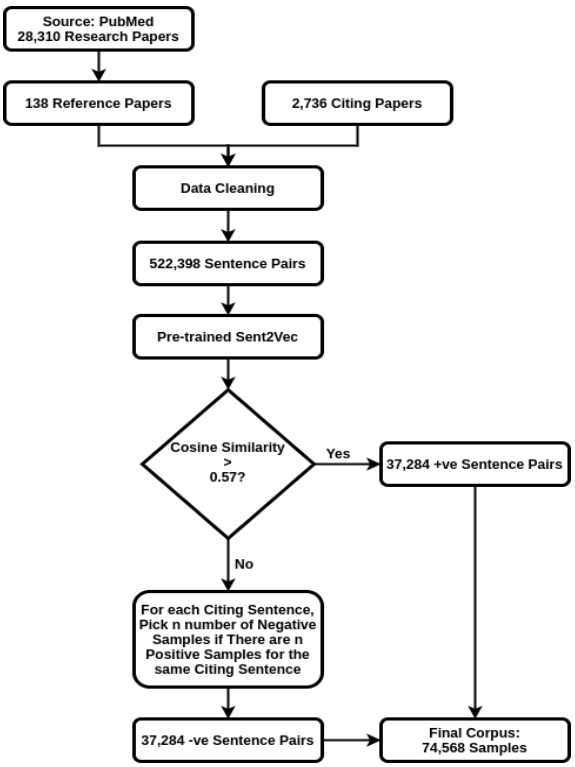

Figure 1: Annotated sentence pair creation for synthetic corpus build-up.

5. Evaluation of the Synthetic Corpus's Effectiveness

We have evaluated the quality of the synthetic corpus in two steps. In the first step, an analysis is performed on a statistically valid sample of the corpus (95% confidence, 3% margin of error) with some human annotators' help, and in the second step, various sequential and tree-structured models are trained with this corpus and the trained models' performances are evaluated on a gold standard test set. For the statistical analysis, from the pool of 74,568 citing and candidate-cited sentence pairs, we randomly selected 750 positive and 750 negative samples for evaluation of the annotation quality (labelled accordingly in the synthetic corpus). Two groups of expert annotators both annotated the 1500 pairs of sentences. There were three people in each group, and they each annotated 500 samples. In other words, each 500-sample chunk was annotated by two people, one from each group. Each reviewer also expressed their level of confidence in the sample annotations they were given. The inter-annotator reliability between the human experts and between the human experts and the synthetic corpus was then calculated using Cohen's κ. One group found 731 positive and 769 negative examples in 1500 sentence pairings, while the other found 709 positive and 791 negative. The annotator groups agreed upon 706 positive samples and 765 negative samples. This study's κ reliability factor is 0.96. For 715 and 701 positive samples, the synthetic silver corpus and the first and second annotator groups agreed on annotation decisions, respectively. In both situations, the annotators agreed with the synthetic silver corpus on all of the negative samples' annotations. In terms of κ, the first group of annotators and the mechanically created corpus have an inter-rater reliability of 0.95 and between the second annotator group and the synthetic corpus, 0.93. By comparing these two sets

Table 4: Performance analysis of different models trained with the gold corpus (Houngbo and Mercer, 2017). The test set contains 400 samples from (Houngbo and Mercer, 2017). The performance metrics are TP: true positive; FP: false positive; TN: true negative; FN: false negative, P: precision, R: recall, F1: F1 score, MCC: Matthews correlation coefficient; Acc: accuracy, BAcc: balanced accuracy.

Model	TP	FP	TN	FN	P	R	F1	MCC	Acc (in %)	BAcc (in %)
hCNN	2	0	390	8	1	0.2	0.33	0.44	98	60
Bi-LSTM & Max-Pooling	1	0	390	9	1	0.1	0.18	0.31	97.75	55
Bi-LSTM & Inner Attention	1	2	398	9	0.33	0.1	0.15	0.17	97.25	54.74
Bi-LSTM & Hierarchical Attention	1	2	398	9	0.33	0.1	0.15	0.17	97.25	54.74
CT-Transformer	2	1	389	8	0.67	0.2	0.31	0.36	97.75	59.87
DT-Transformer	1	2	398	9	0.33	0.1	0.15	0.17	97.25	54.74

Table 5: Performance analysis of different models trained with the synthetic silver corpus. The test set contains 400 samples from (Houngbo and Mercer, 2017). The performance metrics are the same as for Table 4.

Model	TP	FP	TN	FN	P	R	F1	MCC	Acc (in %)	BAcc (in %)
hCNN	7	9	381	3	0.44	0.7	0.54	0.54	97	83.85
Bi-LSTM & Max-Pooling	7	7	383	3	0.5	0.7	0.58	0.58	97.5	84.10
Bi-LSTM & Inner Attention	8	6	384	2	0.57	0.8	0.67	0.67	98	89.23
Bi-LSTM & Hierarchical Attention	8	5	385	2	0.62	0.8	0.69	0.69	98.25	89.35
CT-Transformer	9	5	385	1	0.64	0.9	0.75	0.75	98.5	94.36
DT-Transformer	9	3	387	1	0.75	0.9	0.82	0.82	99	94.62

of results, we can see that the automatic annotations closely match the expert annotations. When evaluating these high κ values, it is important to keep in mind that the annotators were given a 50/50 distribution of positive and negative samples .

For assessing the effectiveness of the introduced silver standard synthetic corpus, we conducted three experiments. In the first experiment, we trained different sequential and tree-structured deep neural network models with 3057 samples with 61 positive samples from the gold standard dataset (Houngbo and Mercer, 2017) and tested them against 400 sentence pairs containing 10 positive sentence pairs from the same dataset. The remaining data from this dataset was used for the validation purpose. In the second experiment, we trained the same models with the synthetic silver standard data and then validated, and tested against the gold standard data just like we did in the first experiment. If the results are found better in the second case, then it proves the effectiveness of training models with the proposed synthetic corpus. In our last experiment, 3057 samples containing 61 positive samples are used for the testing purpose and the remaining data are used for the validation of the models. Results from this experiment shows how good the models perform on a larger portion of the gold standard dataset if they are trained with our synthetic dataset.

The base for all of the models used for the assessment of the quality of the synthetic corpus is the Infersent (Conneau et al., 2017) architecture. As the sentence encoders in the Infersent architecture, four sequential and two tree-based models are used. The basic working principle of Infersent is the use of siamese sentence encoders and applying concatenation, absolute difference, and point wise multiplications over the sentence representations computed from the identical encoders. Finally, this feature representation is used for the downstream tasks. In our experiments, one encoder is fed with the citing sentence and the other encoder is fed with the candidate cited sentence. Then after the encoding and the above stated three operations are done, it is fed to a two-way *softmax* classifier layer for computing the binary semantic relatedness value. Outcome 1 indicates that the citing sentence is actually referring to the candidate cited sentence and 0 indicates the opposite.

In the Infersent architecture, Bi-LSTM with max-pooling, hierarchical CNN (Zhao et al., 2015), Bi-LSTM with inner (Liu et al., 2016) and hierarchical attention (Yang et al., 2016) mechanisms, and two variants of tree-transformers, dependency (DT-Transformer) and constituency (CT-Transformer) tree-transformers (Ahmed et al., 2019), are used as the encoders. All of the encoder architectures are fed with word embeddings from Bio-RoBERTa (Lewis et al., 2020). The hidden layer in all models contains 512 neurons in all cases and a stochastic gradient descent optimizer is used. The hierarchical CNN (hCNN) concatenates features from 4 layers of convolution operations and both the inner and hierarchical attention mechanisms come with 4 heads for focusing on 4 different portions of the sentences which are concatenated

Table 6: Performance analysis of different models trained with the silver standard synthetic corpus. The test set contains 3057 sentence pairs from (Houngbo and Mercer, 2017). The performance metrics are the same as for Table 4.

Model	TP	FP	TN	FN	P	R	F1	MCC	Acc (in %)	BAcc (in %)
hCNN	46	576	2420	15	0.07	0.75	0.13	0.20	80.69	78.09
Bi-LSTM & Max-Pooling	53	359	2637	8	0.13	0.87	0.22	0.31	88.02	87.45
Bi-LSTM & Inner Attention	54	349	2647	7	0.13	0.89	0.23	0.32	88.38	88.43
Bi-LSTM & Hierarchical Attention	56	339	2657	5	0.14	0.92	0.25	0.34	88.75	90.24
CT-Transformer	57	315	2681	4	0.15	0.93	0.26	0.35	89.56	91.46
DT-Transformer	57	301	2695	4	0.16	0.93	0.27	0.36	90.02	91.70

in the end. Both tree-transformers use 6 parallel heads with 50-dimensional key, query and value matrices and the Adagrad optimizer is used. For all of the sentence encoder models, the learning rate is initialized to 0.1. This learning rate is divided by 5 if the validation accuracy reduces in the subsequent epoch.

Tables 4 and 5 show the performances of the models over the same test set containing 400 sentence pairs from Houngbo and Mercer (2017)'s human annotated corpus averaged with four similarly sized randomly chosen subsets. When the models are trained with training set data from the gold standard corpus (3057 samples containing 61 positive samples), no model could retrieve more than 2 out of 10 positive samples from the test set (Table 4). The overall accuracy found for all the models are always more than 97% as the data contains more than 97% negative samples. It proves that when the models are trained with this human annotated corpus, they are biased towards the negative outcome. But, when the same models are trained with the proposed silver standard corpus, the models retrieve 7 to 9 positive samples out of 10 correctly. The best result is found for the DT-Transformer model. It accurately determines 9 positive samples with a balanced accuracy of 94.62%. These results prove the effectiveness of the proposed silver standard dataset.

Table 6 shows the performance of the various models on the original gold standard training set (Houngbo and Mercer, 2017) averaged with four similarly sized randomly chosen subsets when trained with the synthetic silver standard corpus. When the models are trained with the silver corpus, models achieve up to 91.70% balanced accuracy. These models utilize recent deep learning techniques and attention mechanisms which allow them to put more focus on the important portions of the text. The tree-transformer models outperform all the sequential models as they incorporate word level dependency and phrase level information. With these tree structured transformer models, 57 out of 61 positive pairs are extracted accurately. These results reflect that if the models are trained with the proposed synthetic corpus, they perform very well over the gold standard dataset.

6. Conclusion

In this paper, we introduce a synthetic silver standard corpus for the citation linkage task in the biomedical domain and also a method to annotate such a corpus without any human help or expert opinion. Performance of the models trained with this dataset reflects the effectiveness of this corpus. This corpus will be made publicly available. As we started this project a couple of years ago, we used Sent2Vec for the sentence embedding. In future work, different BERT-based models can be utilized. One limitation of this work is that the considered citation text span is limited to a single sentence only. However, in real application scenarios, the referenced text may span over multiple sentences. Keeping this in mind, we are trying to build a gold and a silver standard corpus for the citation linkage task where the text span can be single to multiple sentences.

7. Bibliographical References

Ahmed, M., Samee, M. R., and Mercer, R. E. (2019). You only need attention to traverse trees. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 316–322.

Bonin, S., Petrera, F., Niccolini, B., and Stanta, G. (2003). PCR analysis in archival postmortem tissues. *Molecular Pathology*, 56(3):184–186.

Conneau, A., Kiela, D., Schwenk, H., Barrault, L., and Bordes, A. (2017). Supervised learning of universal sentence representations from natural language inference data. *arXiv preprint arXiv:1705.02364*.

Garfield, E. (1972). Citation analysis as a tool in journal evaluation. *Science*, 178(4060):471–479.

Garzone, M. and Mercer, R. E. (2000). Towards an automated citation classifier. In *Proceedings of the 13th Biennial Conference of the Canadian Society for Computational Studies of Intelligence*, pages 337–346.

Gidiotis, A., Stefanidis, S., and Tsoumakas, G. (2020). Auth@ clscisumm 20, laysumm 20, longsumm 20. In *Proceedings of the First Workshop on Scholarly Document Processing*, pages 251–260.

Houngbo, H. and Mercer, R. E. (2017). Investigating citation linkage with machine learning. In *Proceed-*

ings of the 30th Canadian Conference on Artificial Intelligence, pages 78–83.

Houngbo, K. H. (2017). *Investigating Citation Linkage Between Research Articles*. Ph.D. thesis, The University of Western Ontario.

Huijsmans, C. J., Damen, J., van der Linden, J. C., Savelkoul, P. H., and Hermans, M. H. (2010). Comparative analysis of four methods to extract DNA from paraffin-embedded tissues: Effect on downstream molecular applications. *BMC Research Notes*, 3(1):239.

Kayser, K., Stute, H., Lübcke, J., and Wazinski, U. (1988). Rapid microwave fixation—a comparative morphometric study. *The Histochemical Journal*, 20(6-7):347–352.

Lewis, P., Ott, M., Du, J., and Stoyanov, V. (2020). Pretrained language models for biomedical and clinical tasks: Understanding and extending the state-of-the-art. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 146–157.

Li, L., Zhang, Y., Mao, L., Chi, J., Chen, M., and Huang, Z. (2017). CIST@CLSciSumm-17: Multiple features based citation linkage, classification and summarization. In *BIRNDL 2017*, pages 43–54.

Li, L., Mao, L., Zhang, Y., Chi, J., Huang, T., Cong, X., and Peng, H. (2018). Computational linguistics literature and citations oriented citation linkage, classification and summarization. *International Journal on Digital Libraries*, 19(2-3):173–190.

Li, L., Zhu, Y., Xie, Y., Huang, Z., Liu, W., Li, X., and Liu, Y. (2019). CIST@CLSciSumm-19: Automatic scientific paper summarization with citances and facets. In *BIRNDL 2019*.

Liu, Y., Sun, C., Lin, L., and Wang, X. (2016). Learning natural language inference using bidirectional LSTM model and inner-attention. *arXiv preprint arXiv:1605.09090*.

Mercer, R. (2016). Locating and extracting key components of argumentation from scholarly scientific writing. *Dagstuhl Reports*, 6(4):3–15.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119.

Pagliardini, M., Gupta, P., and Jaggi, M. (2018). Unsupervised learning of sentence embeddings using compositional n-gram features. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 528–540.

Palau, R. M. and Moens, M.-F. (2009). Argumentation mining: The detection, classification and structure of arguments in text. In *Proceedings of the 12th International Conference on Artificial Intelligence and Law*, pages 98–107.

Radev, D. R., Jing, H., and Budzikowska, M. (2000). Centroid-based summarization of multiple documents: Sentence extraction, utility-based evaluation, and user studies. In *NAACL-ANLP 2000 Workshop: Automatic Summarization*.

Ritchie, A., Robertson, S., and Teufel, S. (2008). Comparing citation contexts for information retrieval. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, pages 213–222.

Rose, S., Engel, D., Cramer, N., and Cowley, W. (2010). Automatic keyword extraction from individual documents. *Text Mining: Applications and Theory*, 1:1–20.

Umapathy, A., Radhakrishnan, K., Jain, K., and Singh, R. (2020). CiteQA@CLSciSumm 2020. In *Proceedings of the First Workshop on Scholarly Document Processing*, pages 297–302.

Wang, Y., Carlton, V. E., Karlin-Neumann, G., Sapolsky, R., Zhang, L., Moorhead, M., Wang, Z. C., Richardson, A. L., Warren, R., Walther, A., et al. (2009). High quality copy number and genotype data from FFPE samples using molecular inversion probe (MIP) microarrays. *BMC Medical Genomics*, 2(1):8.

Wolffs, P., Grage, H., Hagberg, O., and Rådström, P. (2004). Impact of DNA polymerases and their buffer systems on quantitative real-time PCR. *Journal of Clinical Microbiology*, 42(1):408–411.

Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., and Hovy, E. (2016). Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489.

Zhao, H., Lu, Z., and Poupart, P. (2015). Self-adaptive hierarchical sentence model. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*, pages 4069–4076.